2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

InfAnFace: Bridging the Infant–Adult Domain Gap in Facial Landmark Estimation in the Wild

Michael Wan*†, Shaotong Zhu†, Lingfei Luan†, Gulati Prateek†, Xiaofei Huang†,
Rebecca Schwartz-Mette‡, Marie Hayes‡, Emily Zimmerman†, Sarah Ostadabbas†§
*Roux Institute, Portland, ME, USA
†Northeastern University, Boston, MA, USA
‡University of Maine, Orono, ME, USA
§Corresponding author: ostadabbas@ece.neu.edu

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956647

***Abstract*—We lay the groundwork for research in the algorithmic comprehension of infant faces, in anticipation of applications from healthcare to psychology, especially in the early prediction of developmental disorders. Specifically, we introduce the first-ever dataset of infant faces annotated with facial landmark coordinates and pose attributes, demonstrate the inadequacies of existing facial landmark estimation algorithms in the infant domain, and train new state-of-the-art models that significantly improve upon those algorithms using domain adaptation techniques. We touch on the closely related task of facial detection for infants, and also on a challenging case study of infrared baby monitor images gathered by our lab as part of in-field research into the aforementioned developmental issues[1]**

***Index Terms*—Facial landmark estimation, domain adaptation, prodromal risk screening, computer vision.**

I. Introduction

The development of effective facial landmark estimation algorithms for infants will open up new avenues of research in healthcare and cognitive psychology. For instance, links between early infant motor or oral function and subsequent developmental disruptions—such as autism spectrum disorder [1], cerebral palsy [2], and pediatric feeding disorders [3]—inspire a tantalizing vision of algorithmic screening or even discovery of prodromal (or pre-symptomatic) risk markers, grounded on automatic facial landmark estimation. However, while vision-based methods have been wildly successful on adult faces, hardly any analogous models exist in the domain of infant faces. Part of the problem is that infants rarely appear in human images "in the wild," and when they do, their faces are often small and obscured and hence not useful for the purposes of facial landmark estimation.

In this paper, we present multi-pronged efforts intended to initiate and foster an ecosystem of research in infant facial landmark estimation. These efforts include:

- the creation of the first-ever **Inf**ant **An**notated **Face**s (InfAnFace) dataset, consisting of 410 images of infant faces with labels for 68 facial landmark locations and various pose attributes (see Fig. 1);
- the organization of InfAnFace into a Train set, intended for machine learning, and an independently-sourced Test set, intended to serve as the field benchmark;

[1]Dataset and model code available at https://github.com/ostadabbas/Infant-Facial-Landmark-Detection-and-Tracking.

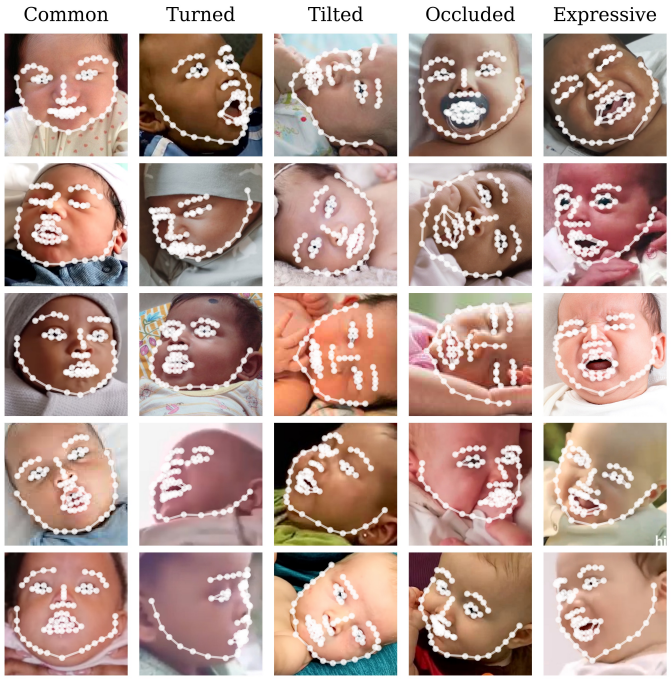

Fig. 1. InfAnFace images and ground truth facial landmarks, grouped by annotated attributes.

- the demonstration of an infant–adult domain gap for existing facial landmark estimation algorithms;
- successful steps towards bridging this gap using domain adaptation techniques under "small-data" conditions, achieving state-of-the-art estimation performance;
- a brief discussion on algorithmic face detection for infants, a prerequisite task for landmark estimation; and
- a report on a challenging real-life application of our algorithms to infrared baby monitor footage captured in-field by our lab.

Our goal is to cast a wide net and encourage further research along each of these lines of inquiry.

II. Related work

Our work relates to a broad swath of prior computer vision research on faces. There are several datasets of faces "in the wild," used to train and benchmark facial landmark estimation algorithms. Datasets featuring 68 facial landmarks adhering to

the Multi-PIE layout [4], which we adopt, include the Helen dataset of 2000 training and 300 test images, the annotated faces in the wild (AFW) dataset of 205 images, featuring large variation in head poses and multiple faces per image [5], and various well-known datasets stemming from the 300 Faces In-the-Wild Challenge (300-W) [6], which we make extensive use of below. Infant faces are hardly represented at all, for instance comprising of only 1.4% of the diverse AFW dataset.

Facial landmark estimation has a long history, with many pioneering models based on fitting face meshes or face models recently surpassed by pure convolutional neural network (CNN) regression methods (see surveys in [7] and [8]). While in principle, face model methods could help with the wide poses inherent in the infant domain, we are only aware of one attempt to build such a model for infants from 3D scans [9], and it is not available for public use. For our study, we stick with flexible and more powerful CNN regression models. We test and modify the well-established high-resolution network (HRNet) model, which uses multi-resolution CNN blocks to carry high-fidelity representations throughout the entire network [10]. We also test with the recent 3FabRec model [11], which achieves few-shot facial landmark localization via unsupervised autoencoder pre-training.

Face detection, a prerequisite for facial landmark estimation, is another mature field, again dominated by CNN methods (see [12] for a survey). We perform tests on infant faces using the recent state-of-the-art RetinaFace model [13]. Finally, recent work by our team tackles the closely-related problem of infant *body pose* estimation, using adversarial domain adaptation methods [14].

III. The InfAnFace Dataset

To establish the first dataset of **In**fant **An**notated **Face**s (InfAnFace), a total of 410 images were sourced and annotated with 68 facial landmarks plus pose attributes, by a team of researchers, to encourage variety and independence. Infants are only featured in a small fraction of human photographs, and when they do appear, their faces are often small or obscured, so gathering a sizable set of images useful for facial landmark estimation algorithms with deep structures is non-trivial. Facial annotations are also more difficult for infants because features are less well-defined, poses are more varied, and natural obstructions like pacifiers are more prevalent. Thus, we believe these diverse sets of well-annotated images represent a hard-earned contribution to a field of study which is itself in its infancy. The results reported later in this paper demonstrate the effectiveness of InfAnFace for machine learning training and testing.

Landmark annotations. The 68 facial landmark annotations adhere to the industry standard Multi-PIE layout [4]. We have also included coordinates of the *minimal bounding box*, the smallest upright box containing all of the landmarks. Following industry conventions for 2D-based annotations, landmarks obscured by the face itself (e.g. when turned) were assigned to the nearest point on the face which *is* visible in the image, as the true projected position is hard to estimate. We employed both a standard annotation tool [15], and the specialized human–artificial intelligence hybrid tool AH-CoLT [16], with the landmark predictions of the FAN model [17] serving as the starting point for the human annotations.

Face attributes. To facilitate analysis, we included binary annotations for each image, indicating whether infant's face is: *turned* (if the eyes, nose, and mouth are not clearly visible), *tilted* (if the head axis, projected on the image plane, is 45° or more beyond upright), *occluded* (if landmarks are covered by body parts or objects), and excessively *expressive* (if the facial muscles are tense, as when crying, laughing, etc.).

Sourcing and the Test–Train split. InfAnFace images were sourced from Google Images and YouTube via a wide range of search queries to obtain a diversity of appearances, poses, expressions, scene conditions, and image quality. Approximately two-fifths of the infants represented appear to be non-white. For our machine learning experiments, we split InfAnFace into dedicated training and test sets, with the following composition: **InfAnFace Train**, consisting of 210 images, with 51- and 55-image batches drawn respectively from Google and YouTube, and a 105-image batch drawn from a specialized search on YouTube for infant formula advertisements; and **InfAnFace Test**, consisting of 200 images, with two 100-image batches drawn respectively from Google and YouTube. These five "batches" were sourced and annotated by rotating configurations of researchers, ensuring a level of variety and independence between them and hence between InfAnFace Train and InfAnFace Test as well.

Attributes and the Common–Challenging split. The breakdown of annotated attributes across InfAnFace and its Train and Test subsets, shown in Table I, demonstrates that all three sets exhibit a healthy diversity of ideal and adverse conditions. Each attribute is present in each set to a large enough degree to be useful for training or analysis, but the distributional differences between InfAnFace Train and Test also highlight their independence. To aid interpretability of experiments performed on InfAnFace Test, we partition it further into two subsets based on the annotated attributes: **InfAnFace Common**, the subset 80 images with faces free from all adverse attributes; and **InfAnFace Challenging**, the remaining set of 120 images exhibiting at least one of them.

To sum up, InfAnFace (410 images) is split into InfAnFace Train (210) and InfAnFace Test (200), and InfAnFace Test into InfAnFace Common (80) and InfAnFace Challenging (120).

IV. Illustrating The Infant-Adult Domain Gap

We examine the facial landmark domain gap between infants and adults by comparing InfAnFace Test with three predominantly adult image sets: 300-W Test (600 images),

TABLE I
Attribute incidence rates in InfAnFace subsets

Attribute	% of Inf. Train	% of Inf. Test	% of InfAnFace
Turned	28.6	12.5	20.2
Tilted	28.6	36.5	32.0
Occluded	26.2	14.0	20.2
Expressive	37.1	14.5	26.1

300-W Common (554 images), and 300-W Challenging (135 images) [6]. Our definitions for InfAnFace Test, Common, and Challenging loosely track with these 300-W sets, with Common featuring relatively easy poses, Test a balanced mix of poses, and Challenging the most difficult poses. The 300-W images are partially taken from earlier datasets so a range of sources is represented. We start with a brief comparison of geometric attributes, and then turn to a study of facial landmark estimation model performance on InfAnFace vs. 300-W, before capping off the section with a visualization of one of the model's internal representations of each image.

A. Geometric observations

We record some geometric measures of interest across various adult and infant subsets in Table II. The first column shows that the aspect ratios of the minimal bounding boxes for mean adult faces are $\sim 1:1$, compared to $\sim 5:4$ for infants. The second column of Table II records the mean ratios of two commonly used normalization factors in facial landmark estimation. We will discuss these normalization factors and the significance of these values in the next section. See Supp. Section A-A for a visual comparison of facial geometries.

B. Benchmarking facial landmark estimation

In order the demonstrate the gap in facial landmark estimation between adults and infants, and to establish baseline performance metrics, we performed predictions using two recent 2D facial landmarking models, HRNet [10] (with the pretrained HRNetV2-W18 model) and 3FabRec [11] (with the pretrained `lms_300w` model). A selection of predicted landmarks is shown in Fig. 2, which also includes improved predictions from our own model (described in Section V).

The main error metric for facial landmark estimation is the *normalized mean error (NME)*: the mean Euclidean distance between each predicted landmark and the corresponding ground truth landmark, divided by a normalization factor with the same units (pixels), to achieve scale-independence. We consider two normalization factors: the *interocular distance (iod)*, defined as the distance between the two outer corners of the eyes, and the *minimal bounding box size (box)*, defined as the geometric mean of its height and width. The means of the ratios of these two normalization factors across various infant and adult sets is recorded in the second column of Table II. Those values show that the gap between metric performance on adults and infants will be greater on average under the box norm, compared to the interocular norm. Neither error metric is likely to be domain invariant, so using both provides a more rounded comparison. Note that although NME is sometimes reported as a percentage, it can exceed 100 in value.

TABLE II
GEOMETRIC VALUES ACROSS ADULT 300-W AND INFANFACE IMAGES

Image Set	$\frac{\mu(\text{min. box width})}{\mu(\text{min. box height})}$	$\mu\left(\frac{\text{min. box size}}{\text{interocular dist.}}\right)$
300-W Test	1.01	1.69
300-W Common	1.03	1.67
300-W Challenging	1.02	1.85
InfAnFace Test	1.28	1.53
InfAnFace Common	1.28	1.48
InfAnFace Challenging	1.27	1.56

TABLE III
LANDMARK ESTIMATION NORMALIZED MEAN ERROR (NME, ↓ IS BETTER) UNDER iod AND box NORMALIZATIONS, ON ADULT 300-W AND INFANFACE IMAGES

	300-W Test		300-W Comm.		300-W Chal.	
Model	iod ↓	box ↓	iod ↓	box ↓	iod ↓	box ↓
3FabRec (*CVPR* '20)	3.85	2.24	3.40	2.04	5.74	3.10
HRNet (*TPAMI* '21)	3.85	2.28	2.92	1.76	5.13	2.77
	InfAnFace Test		InfAnFace Comm.		InfAnFace Chal.	
Model	iod ↓	box ↓	iod ↓	box ↓	iod ↓	box ↓
3FabRec (*CVPR* '20)	12.59	8.13	*4.93*	*3.36*	17.69	11.31
3FabRec-JT	9.28	6.04	**4.90**	**3.34**	12.20	7.84
3FabRec-FT	9.00	5.86	5.30	3.59	11.47	7.37
3FabRec-R90	*8.45*	*5.52*	5.45	3.71	*10.45*	*6.73*
3FabRec-R90JT	**8.00**	**5.22**	5.12	3.49	**9.93**	**6.39**
HRNet (*TPAMI* '21)	12.82	8.35	5.07	3.45	17.98	11.62
HRNet-JT	5.95	3.87	**4.47**	**3.03**	6.94	4.43
HRNet-FT	5.90	3.85	*4.48*	*3.04*	6.84	4.38
HRNet-R90	5.90	3.87	4.88	3.31	6.58	4.24
HRNet-R90JT (ours)	**5.30**	**3.47**	4.52	3.07	**5.82**	**3.73**
HRNet-R90FT (ours)	*5.40*	*3.52*	4.68	3.17	*5.88*	*3.78*

TABLE IV
LANDMARK ESTIMATION FAILURE RATE (FR, OUT OF 100, ↓ IS BETTER) AND AREA UNDER THE CURVE (AUC, OUT OF 100, ↑ IS BETTER) AT $\text{NME}_{\text{iod}} = 10$, ON ADULT 300-W AND INFANFACE IMAGES

	300-W Test		300-W Comm.		300-W Chal.	
Model	FR ↓	AUC ↑	FR ↓	AUC ↑	FR ↓	AUC ↑
3FabRec (*CVPR* '20)	1.17	52.72	0.00	64.73	2.22	38.95
HRNet (*TPAMI* '21)	0.33	61.54	0.18	70.83	2.22	48.84
	InfAnFace Test		InfAnFace Comm.		InfAnFace Chal.	
Model	FR ↓	AUC ↑	FR ↓	AUC ↑	FR ↓	AUC ↑
3FabRec (*CVPR* '20)	21.00	*36.66*	*0.00*	**51.69**	35.00	26.63
3FabRec-JT	17.00	**37.34**	1.25	*51.38*	27.50	**27.98**
3FabRec-FT	*14.50*	35.29	*0.00*	46.88	*24.17*	*27.57*
3FabRec-R90	18.00	33.21	1.25	46.15	29.17	24.58
3FabRec-R90JT	**14.00**	36.16	1.25	49.43	**22.50**	27.32
HRNet (*TPAMI* '21)	22.50	34.75	0.00	49.30	37.50	25.06
HRNet-JT	7.00	45.52	0.00	**55.32**	11.67	38.99
HRNet-FT	6.00	45.91	0.00	*55.22*	10.00	39.71
HRNet-R90	5.00	43.75	0.00	51.18	8.33	38.80
HRNet-R90JT (ours)	**2.50**	**48.07**	0.00	54.69	**4.17**	**43.65**
HRNet-R90FT (ours)	*3.00*	*46.87*	0.00	53.23	*5.00*	*42.63*

We tabulate the mean NMEs under both normalizations and of both model's landmark predictions across various adult and infant image sets in Table III. Further characterizations of landmark estimation performance include the *failure rate (FR)* of images in the dataset with NME greater than a set threshold, and the *area under the curve (AUC)* of the cumulative NME distribution up to a set threshold. Table IV reports the FR and AUC for NME normalized by the interocular distance, with a threshold of $\text{NME} = 10$. (Both tables also include results from our improved models, to be discussed in Section V.) Fig. 3 plots the cumulative NME distribution curves themselves,

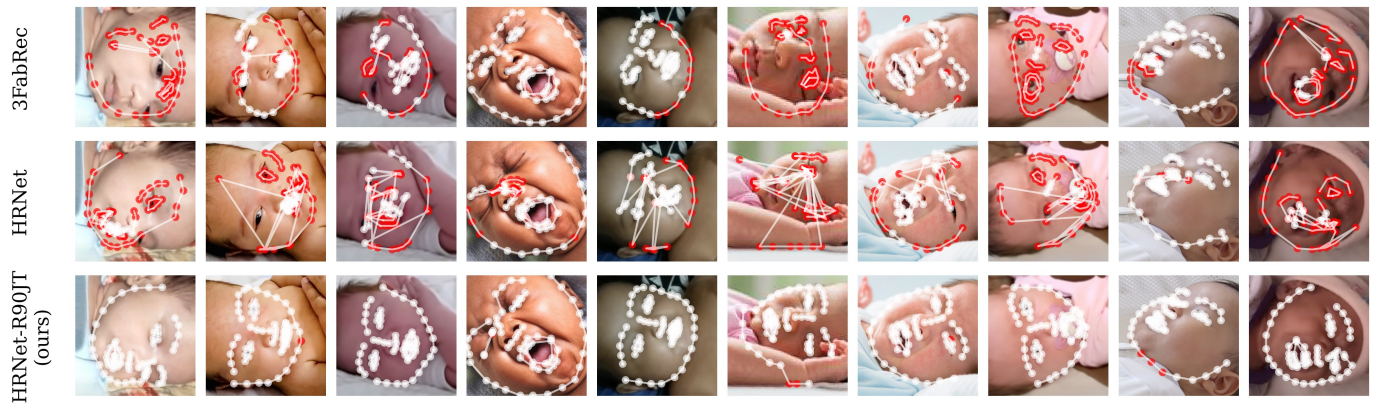

Fig. 2. Model predictions on InfAnFace Test images, from 3FabRec, HRNet, and our state-of-the-art HRNet-R90JT model. Landmark predictions with high error are highlighted in red.

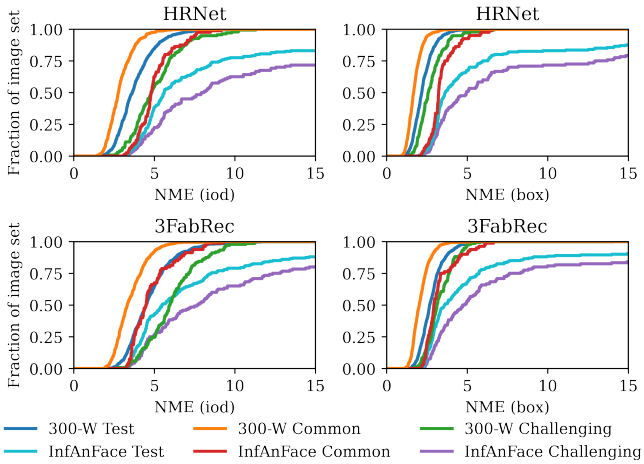

Fig. 3. Cumulative normalized mean error (NME) curves for HRNet and 3FabRec facial landmark estimation models, highlighting the performance gap between infants (InfAnFace) and adults (300-W). Errors under both the interocular (iod) and minimal bounding box (box) normalization factors shown.

under both normalizations, and with a threshold of NME $= 15$ for greater context.

These results show a significant performance gap between adult and infant domains as a whole, which in line with our expectations from Table II is more pronounced under the box norm. Within each domain, performance degrades from the Common to Test to Challenging sets. Furthermore, poor performance on InfAnFace Test seems largely attributable to the difficulty of the InfAnFace Challenging subset. These considerations, as well as a visual inspection of landmark predictions as in Fig. 2, expose adverse conditions defining the InfAnFace Challenging subset, such as tilt and occlusion, as likely causes of poor facial landmark estimations on infant faces. We believe, though, that such conditions are endemic to infant images captured in the wild, and thus, infant-focused algorithms should seek to overcome them.

We note that while the performance results on InfAnFace Challenging are the most dramatic, the tables and graphs do also reveal a smaller but definite performance gap between results on InfAnFace Common and the adult image sets, particularly the 300-W Common and 300-W Test sets which

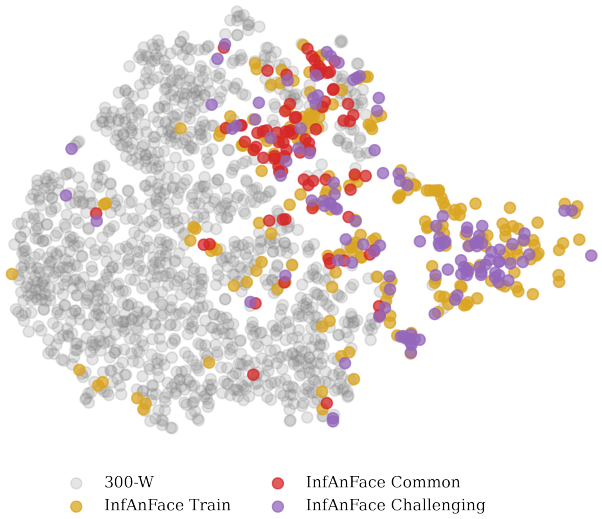

Fig. 4. A t-SNE visualization of the internal HRNet representations of adult 300-W and InfAnFace images, highlighting the domain gap between them. InfAnFace Common is more closely integrated with the adult images compared to the more divergent InfAnFace Challenging. Together they comprise InfAnFace Test, which has roughly the same distribution as InfAnFace Train.

arguably serve as fairer points of comparison. This suggests that even in more ideal pose conditions, a domain gap exists.

C. t-SNE visualization

An elegant visual companion to the preceding analysis can be found in the t-distributed stochastic neighbor embedding (t-SNE) plots in Fig. 4, which offer a glimpse into the how the HRNet neural network "perceives" each image relative to one another. Each infant or adult image is processed by HRNet into a $270 \times 64 \times 64$-dimensional vector (before the regression head), and the set of these representations is compressed by the t-SNE algorithm [18] into a set of two-dimensional coordinates for each image, with relative relationships probabilistically preserved. This set of coordinates is plotted in Fig. 4, with image set membership highlighted by color[2]. In line with our earlier observations, the t-SNE distribution highlights a general

[2]We employed the scikit-learn [19] implementation of t-SNE, with perplexity 50 and otherwise default settings.

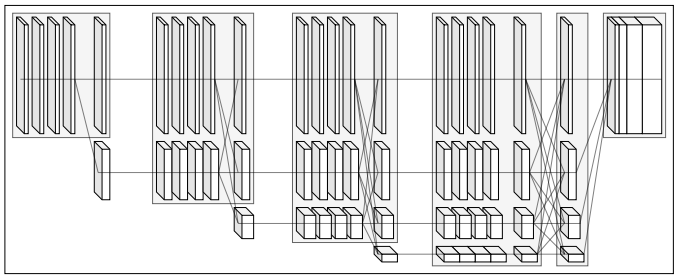

Fig. 5. Structure of high-resolution net (HRNet) [10], the backbone for our best models, featuring parallel multi-resolution convolutional layers.

gap between adult and infant image sets, and within InfAnFace Test, a split between the tamer InfAnFace Common subset and the more divergent InfAnFace Challenging subset. InfAnFace Test as a whole is distributed similarly to InfAnFace Train.

V. Bridging The Infant-Adult Domain Gap

We now turn to our efforts to solve facial landmark estimation for infant faces. Since there exist strong models and ample data in the adult domain, for this inceptive paper we apply the domain adaptation tools of joint-training and fine-tuning, in addition to targeted data augmentation tools based on insights from our analysis of the domain gap. We have already exhibited the effectiveness of our resulting algorithm in Fig. 2, and our quantitative analysis will show that, for instance, our modified HRNet model performs with a failure rate of 2.50% at $\text{NME}_{\text{iod}} = 10$ on InfAnFace Test, drastically improving upon the 22.50% for the original HRNet model, and even approaching the 1.17% failure rate of the original HRNet on the adult 300-W Test set. Note that commonly used adversarial discriminative domain adaptation methods like [20] or [21] work well for classification but are typically ineffective in the regression domain adaptation setting, where the covariate shift assumption usually holds (as discussed in [22] and especially [23]). So we believe that our methods yield results competitive with what can be obtained from general domain adaptation tools applicable in the landmark estimation setting, and serve as strong baselines for the nascent field.

A. Training implementations

We started with two powerful state-of-the-art models, the high-resolution net (HRNet) [10] which runs multi-resolution convolutional layers connected in parallel—see Fig. 5 for a network diagram—and 3FabRec [11], which features a autoencoder ResNet structure that learns facial representations in an unsupervised manner before switching to supervised learning for facial landmark estimation.

For HRNet, we performed extensive validation set experiments (not involving our final test data) to test a number of joint-training and fine-tuning paradigms, and a range of training data augmentations including random zooms and rotations, inspired by our earlier observations that landmark estimation is often poor for infants with difficult poses. Our tests yielded two final models with similar validation performance:

HRNet-R90JT, HRNet **j**ointly **t**rained on the union of the 300-W training set and InfAnFace Train, with **r**otations of angles between $\pm\mathbf{90}^\circ$ applied to the training data 3/5ths of the time.

HRNet-R90FT, where HRNet is first trained on 300-W data, then **f**ine-**t**uned by further training on InfAnFace Train with a reduced learning rate and parameters frozen before the second HRNet layer, and with **r**otations of angles between $\pm\mathbf{90}^\circ$ applied 3/5ths of the time for both phases of training.

The hyperparameters and configurations that were chosen by validation included:

1) the rotation augmentation range of $\pm 90^\circ$, from the choices $\{\pm 30^\circ, \pm 60^\circ, \pm 90^\circ, \pm 120^\circ, \pm 150^\circ\}$;
2) the random zoom range of $100\% \pm 25\%$, from the choices $\{\pm 5\%, \pm 10\%, \pm 15\%, \pm 20\%, \pm 30\%, \pm 50\%\}$;
3) for fine-tuning runs, the starting learning rate of 10^{-7}, from the choices $\{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$;
4) for fine-tuning runs, the choice to freeze the model before the second HRNet layer, from choice of second, third, or fourth layers, or leaving the entire model unfrozen (chosen simultaneously with the previous parameter in an exhaustive grid search); and
5) the choice of which epoch's checkpoint to take to represent the model, upon conclusion of a training run.

To demonstrate the necessity of the features employed in our best models above, we also trained three further HRNet models with some of these features selective ablated:

HRNet-JT, HRNet **j**ointly **t**rained on the union of the 300-W training set and InfAnFace Train, with HRNet's default rotations of angles between $\pm 30^\circ$ applied to the training data 3/5ths of the time.

HRNet-FT, where HRNet is first trained on 300-W data, then **f**ine-**t**uned by further training on InfAnFace Train with a reduced learning rate and parameters frozen before the second HRNet layer, and with HRNet's default rotations of angles between $\pm 30^\circ$ applied to the training data 3/5ths of the time.

HRNet-R90, HRNet trained only on the 300-W training set, with **r**otations of angles between $\pm\mathbf{90}^\circ$ applied to the training data 3/5ths of the time.

Furthermore, to demonstrate the effectiveness of our technique more generally, we also applied these modifications to the 3FabRec model at the supervised training stage, obtaining:

3FabRec-JT, 3FabRec **j**ointly **t**rained on the union of the 300-W training set and InfAnFace Train.

3FabRec-FT, where 3FabRec is first trained on 300-W data, then **f**ine-**t**uned by further training on InfAnFace Train.

3FabRec-R90, 3FabRec trained only on the 300-W training set, with **r**otations of angles between $\pm\mathbf{90}^\circ$ randomly applied.

3FabRec-R90JT, 3FabRec **j**ointly **t**rained on the union of the 300-W training set and InfAnFace Train, with **r**otations of angles between $\pm\mathbf{90}^\circ$ randomly applied.

These configurations were chosen to mirror our HRNet configurations, rather than by validation hyperparameter search, to allow for direct comparisons.

B. Experimental results and analysis

The infant facial landmark estimation performance metrics for 3FabRec, HRNet, and the modified versions of those

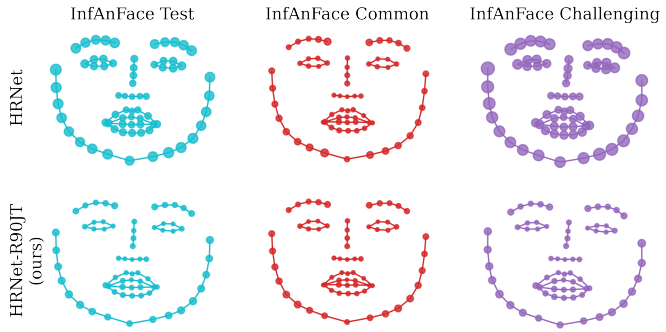

Fig. 6. A visualization depicting the performance per landmark of HRNet against our HRNet-R90JT across various InfAnFace subset, with each landmark radius drawn in proportion to the subset-mean interocular-normalized error for that landmark. (Landmark positions are as in Supp. Fig. S1.)

models are reported in Table III and Table IV, and sample landmark predictions for our best-performing HRNet-R90JT model can be found in Fig. 2. These results show significantly improved performance from our HRNet-R90JT and HRNet-R90FT models compared to the original 3FabRec and HRNet, bringing infant facial landmark estimation results within striking range of the best results on adult faces. The most notable gains come from performance on InfAnFace Challenge, but there are also small gains in performance on InfAnFace Common. A visualization of the the marked improvements of HRNet-R90JT over HRNet on a per-landmark level can be found in Fig. 6, and overall cumulative error curves can be found in Supp. Section A-B.

The reported results on the models with ablated features, HRNet-JT, HRNet-FT, and HRNet-R90, demonstrate that both the infant training data and the wider rotation angles are needed to realize the full gains. Intriguingly, though, either feature on its own seems capable of taming the most egregious predictions on InfAnFace Challenge. We also observe that the effect of applying wider rotation angles in training, on models trained on infant data, is to notably improve performance on InfAnFace Challenging faces at only a slight cost of performance on the largely-upright faces in InfAnFace Common.

The performance of the modified 3FabRec models are weaker than those of correspondingly modified modified HRNet models, perhaps in part due to the exhaustive hyperparameter validation tuning performed for the latter. However, the relative performances of the different 3FabRec variants support our general conclusions that while domain adaptation and data augmentation are both individually quite effective for improving performance, their combination leads to further gains; and that most gains come from improvements to the difficult InfAnFace Challenging images.

VI. Infant Face Detection

Facial landmark estimation algorithms usually require information about face location as input, typically in the form of coordinates for a bounding box. These coordinates are included with test datasets like InfAnFace Test, but for most real-world applications, they need to be obtained beforehand from dedicated *face detection* algorithms. While in principle, such algorithms could be fine-tuned with infant data, we found them to be already adequate out-of-the-box.

To demonstrate this, we applied a cutting-edge face detection algorithm, RetinaFace [13], to InfAnFace Test. For face detection, the main performance metric on a set of predictions is the *average precision*, defined as an interpolated area under an associated precision-recall curve (see Supp. Section A-C). The average precisions of RetinaFace across the InfAnFace Test, Common, and Challenging sets respectively are 98.1%, 100.0%, and 94.7%, attesting to excellent performance on infants. Thus, the combination of RetinaFace and our HRNet-R90FT and HRNet-R90JT models provide complete solutions to the facial landmark estimation task for infants.

VII. An In-Field Baby Monitor Challenge

As part of an ongoing study exploring links between infant motor function and developmental disruptions, our lab gathered baby monitor footage of 15 real infant subjects under an Institutional Review Board (IRB #17-08-19) approval. The videos feature infants napping and sleeping in their own crib or bassinet, typically under low-light conditions, triggering the baby monitor's monochromatic infrared capture mode. These pose and lighting conditions entail significant challenges to algorithmic comprehension.

We annotated 213 private images from these videos in order to gauge facial landmark estimation performance in extremely adverse vision conditions. The predictions of the original HRNet and of our improved HRNet-R90JT on this set yielded respective NMEs of 32.3 and 15.5 under the minimal box size normalization[3], poor in comparison to HRNet-R90JT's error of 3.73 on the InfAnFace Challenging set.

We attempted to mitigate this performance gap with further data augmentation techniques. We established a new model, **HRNet-R150GJT**, **j**ointly **t**rained on 300-W and InfAnFace Train data, with **r**otation augmentations in the range of $\pm$**150**°, and a **g**rayscale filter applied 1/2 of the time, where both hyperparameters were chosen based on validation on InfAnFace Train. The resulting predictions yielded an average NME of 11.7 under the box normalization factor. This is a notable improvement, but deeper developments are needed to raise performance to a higher level.

VIII. Conclusion and Future Work

We have advanced the algorithmic comprehension of infant faces by introducing the first dataset in the field, demonstrating the inadequacies of existing facial landmark estimation algorithms on the infant domain, and training best-in-class models using domain adaptation techniques and our newly introduced training data. We also demonstrated the effectiveness of existing face detection algorithms on infants, yielding a full-stack solution to the landmark estimation problem for infants. Finally, we explored applications to challenging in-field data, and pointed out the need for further research in this area.

[3]Note: The interocular norm is not suitable in this setting because many in-bed infants have their heads turned, hiding one eye, and the industry convention of assigning that eye's landmark to the nearest visible spot on the head renders the interocular distance somewhat arbitrary.

REFERENCES

[1] Jannath Begum Ali, Tony Charman, Mark H Johnson, and Emily JH Jones, "Early motor differences in infants at elevated likelihood of autism spectrum disorder and/or attention deficit hyperactivity disorder," *Journal of autism and developmental disorders*, vol. 50, no. 12, pp. 4367–4384, 2020.

[2] Centers for Disease Control and Prevention, "Screening and diagnosis of cerebral palsy," 2012.

[3] Lene Lindberg, Gunilla Bohlin, and Berit Hagekull, "Early feeding problems in a normal population," *International Journal of Eating Disorders*, vol. 10, no. 4, pp. 395–405, 1991.

[4] Ralph Gross, Iain Matthews, Jeff Cohn, Takeo Kanade, and Simon Baker, "Multi-PIE," *Proceedings of the ... International Conference on Automatic Face and Gesture Recognition. IEEE International Conference on Automatic Face & Gesture Recognition*, vol. 28, no. 5, pp. 807–813, May 2010.

[5] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2879–2886.

[6] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 Faces In-The-Wild Challenge: database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, Mar. 2016.

[7] Xin Jin and Xiaoyang Tan, "Face alignment in-the-wild: A Survey," *Computer Vision and Image Understanding*, vol. 162, pp. 1–22, Sept. 2017.

[8] Yue Wu and Qiang Ji, "Facial Landmark Detection: A Literature Survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, Feb. 2019.

[9] Araceli Morales, Antonio R. Porras, Liyun Tu, Marius George Linguraru, Gemma Piella, and Federico M. Sukno, "Spectral Correspondence Framework for Building a 3D Baby Face Model," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, Buenos Aires, Argentina, Nov. 2020, pp. 708–715, IEEE.

[10] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[11] Bjorn Browatzki and Christian Wallraven, "3FabRec: Fast Few-Shot Face Alignment by Reconstruction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, June 2020, pp. 6109–6119, IEEE.

[12] Shervin Minaee, Ping Luo, Zhe Lin, and Kevin Bowyer, "Going Deeper Into Face Detection: A Survey," *arXiv:2103.14983 [cs]*, Apr. 2021, arXiv: 2103.14983.

[13] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *CoRR*, vol. abs/1905.00641, 2019.

[14] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas, "Invariant representation learning for infant pose estimation with small data," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2021*, December 2021.

[15] Abhishek Dutta and Andrew Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019, MM '19, ACM.

[16] Xiaofei Huang, Behnaz Rezaei, and Sarah Ostadabbas, "AH-CoLT: an AI-Human Co-Labeling Toolbox to Augment Efficient Groundtruth Generation," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2019, pp. 1–6.

[17] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1021–1030, Oct. 2017.

[18] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov 2008.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-Adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, pp. 1–35, May 2016.

[21] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," *CoRR*, vol. abs/1702.05464, 2017.

[22] Corinna Cortes and Mehryar Mohri, "Domain Adaptation in Regression," in *Algorithmic Learning Theory*, Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, Eds., vol. 6925, pp. 308–323. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, Series Title: Lecture Notes in Computer Science.

[23] Antoine de Mathelin, Guillaume Richard, Mathilde Mougeot, and Nicolas Vayatis, "Adversarial weighting for domain adaptation in regression," *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 49–56, 2021.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.